Scientific and Technical Advisory Committee Peer Review for the James River Chlorophyll-*a* Criteria Re-evaluation

Lora Harris[1], Tom Fisher[2], Jim Hagy[3], Dong Liang[1], Martha Sutula[4]

[1]University of Maryland Center for Environmental Science – Chesapeake Biological Laboratory, [2]University of Maryland Center for Environmental Science – Horn Point Laboratory, [3]US EPA, [4]Southern California Coastal Water Research Project

STAC Review Report
October 2016

STAC Publication 16-007

About the Scientific and Technical Advisory Committee

The Scientific and Technical Advisory Committee (STAC) provides scientific and technical guidance to the Chesapeake Bay Program (CBP) on measures to restore and protect the Chesapeake Bay. Since its creation in December 1984, STAC has worked to enhance scientific communication and outreach throughout the Chesapeake Bay Watershed and beyond. STAC provides scientific and technical advice in various ways, including (1) technical reports and papers, (2) discussion groups, (3) assistance in organizing merit reviews of CBP programs and projects, (4) technical workshops, and (5) interaction between STAC members and the CBP. Through professional and academic contacts and organizational networks of its members, STAC ensures close cooperation among and between the various research institutions and management agencies represented in the Watershed. For additional information about STAC, please visit the STAC website at www.chesapeake.org/stac.

Publication Date: October 2016

Publication Number: 16-007

Suggested Citation:

Harris, L., T. Fisher, J. Hagy, D. Liang, M. Sutula. 2016. Scientific and Technical Advisory Committee Peer Review for the James River Chlorophyll-*a* Criteria Re-evaluation. STAC Publication Number 16-007, Edgewater, MD. 41 pp.

Cover graphic from: https://en.wikipedia.org/wiki/James_River

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

The enclosed material represents the professional recommendations and expert opinion of individuals undertaking a workshop, review, forum, conference, or other activity on a topic or theme that STAC considered an important issue to the goals of the CBP. The content therefore reflects the views of the experts convened through the STAC-sponsored or co-sponsored activity.

STAC Administrative Support Provided by:

Chesapeake Research Consortium, Inc.
645 Contees Wharf Road
Edgewater, MD 21037
Telephone: 410-798-1283
Fax: 410-798-0816
http://www.chesapeake.org

Table of Contents

Peer Review for the James River Chlorophyll-*a* Criteria Re-evaluation

This report documents the conclusions reached by a peer review panel convened by the Chesapeake Bay Program's Scientific and Technical Advisory Committee (STAC) to review two reports related to the state of Virginia's numeric water quality criteria for chlorophyll-*a* in the tidal portion of the James River, a tributary of the Chesapeake Bay. The panel was principally charged with reviewing two reports (Robertson 2016 and VA DEQ 2016a) that were developed in parallel for the Virginia Department of Environmental Quality (VA DEQ). VA DEQ (2016a) is the primary document reviewed by this panel and focused on determining whether existing chlorophyll-*a* criteria are protective of designated uses. Robertson (2016) describes a new assessment methodology that the state of Virginia could use to evaluate attainment of its chlorophyll-*a* criteria.

Our charge from STAC was: 1) to provide general feedback on the content, structure, and editorial quality in these reports, and especially whether they clearly convey the information needed to understand and evaluate the scientific arguments presented; 2) to respond to a series of seven questions laid out in the formal review request; and 3) to provide feedback on the chlorophyll-*a* criteria broadly in the context of the Chesapeake Bay ecosystem. Due to the tight timeline for the re-evaluation process, our comments in this interim report are focused on general feedback on the writing (1) and responses to the seven questions (2).

Our panel was composed of a team with particular expertise in Chesapeake Bay estuarine ecology, spatial statistics, and members familiar with development of water quality criteria within a TMDL context. Three members are academics at host institutions in the Chesapeake Bay (Harris, Fisher, Liang), and two member scientists engage in federal (Hagy) and state (Sutula) water research and management efforts. Three members have engaged in recent chlorophyll criteria development efforts both in the Chesapeake Bay (Fisher) and in other national estuaries (Hagy and Sutula). Liang is a statistician with a particular expertise in spatial datasets and methods. Our process centered on assigning lead experts to each of the seven questions, with secondary reviewers engaged in first drafts before distribution to the entire group. We then held two critical conference calls to come to consensus on major comments and recommendations. The report was pulled together for review by the final panel, and consensus recommendations and conclusions were identified. The contents of this report have been reviewed and approved by the entire panel.

Executive Summary

Our Executive Summary (ES) is written to highlight consensus points detailed in the remainder of the report. The comments in this ES are provided as "Responses" to the originally posed questions (provided in Appendix A) and are organized by Question number. Following the Response sections, we provide two final sections discussing "Near Term Recommendations" and a "Longer-Term Recommendation." With focused effort, we believe that the near term recommendations related to the harmful-effects analyses (VA DEQ 2016a) may be accomplished in a period of 6 months to a year, assuming that a full time statistician with experience analyzing environmental data is available for the effort. Coordination of decisions regarding risk thresholds between a scientific team refining the analyses and policy makers also seems feasible on this time table. Refinement of the assessment approach (Robertson 2016) may require iterative analyses that would be best accomplished in parallel to the effects-based analyses in order to satisfy the spatio-temporal issues we identify herein. Configuring a monitoring program to better complement these approaches is a longer term endeavor.

Whether our recommendations will create large changes in magnitude of the outcomes reported by VA DEQ (2016a) or Robertson (2016) is unclear. The temporal scales of aggregation are 'apples and oranges' – the bias issue needs to be resolved, and whether models can (or should) be fit to the data between harmful effects and chlorophyll-*a* will rest on a re-framing of quantitative metrics that are not easily predicted at this time. The panel believes, however, that proceeding with these recommendations will result in criteria that are easier to justify to the public with a more straightforward path to assess waterbody status.

The general approaches we reviewed have redeeming qualities. If the goal is to keep current criteria but provide improved rationale for why they are protective, then VA DEQ (2016a) has made an attempt at this goal. However, to meet a higher standard of understanding that will allow us to quantitatively evaluate what levels of chlorophyll-*a* are truly protective, we encourage the VA DEQ and their designated Scientific Advisory Panel (SAP) to follow our near term recommendations. Finally, VA DEQ (2016a) and Robertson (2016) document appropriate approaches to refine the scientific basis for the chlorophyll-*a* criteria, but the details in the implementation of these analyses are problematic and need to be refined to better support the James River Chlorophyll Criteria (JRCC) and assessment of attainment of those criteria. Outlined below are our specific conclusions and recommendations that summarize our response to each of the formal charge questions. Our responses here follow the order of the questions provided, even as we re-order our full response in the remainder of the document to allow for a more logical connection of the thoughts and concerns described herein.

Response to Question #1:

1. Spatial misalignment between the monitoring data used in the harmful-effects analyses introduce bias into the expected frequency calculations. In particular, such misalignment can cause underestimation of the expected frequency of harmful algal blooms (HABs) when conditioned on mean chlorophyll-*a* concentration.
2. Deriving model-based estimates of standard error is also recommended.
3. We provide recommendations for monitoring to more effectively link datasets to the effects-based analyses.

Response to Question #2:

1. Applying a harmful-effects approach has merit, with ample precedent for the use of such risk-based approaches in establishing the scientific basis for water quality criteria to protect human and ecological health.
2. The reference-based approach is also valuable toward understanding what is achievable in a minimally disturbed condition. The combination of effects-based and reference-based approaches can be used in a complementary fashion to provide policymakers better context to understand how risk changes as a function of increasing stress (chlorophyll-*a*).
3. One critique of the harmful-effects analysis used by the SAP is there are no clear statements relating the policy framework to selection of "low risk" (*i.e.*, protective) versus "high risk" (*i.e.*, non-protective; see Response 3) for levels of harmful effects. Once identified, these quantitative thresholds regarding risks associated with designated uses can then be modeled statistically to identify the chlorophyll-*a* values that will represent these risk categories. A table is recommended to list the various metrics used (HAB toxins, low dissolved oxygen, pH, etc.) with associated thresholds, and identifying where these thresholds intersect with both scientific rationale and policy. Such a table would help to organize the framework.
4. A second critique is that the analysis presented in VA DEQ (2016a) did not clearly articulate the spatial and temporal scales associated with risk of adverse effects and how these were considered in the aggregated data used in the x- and y-axis of the conditional probability analysis. It is important to justify these decisions and show how they link to the extent, frequency and duration specified in the final criteria.

Response to Questions #3 and #4:

1. The current terminology used in VA DEQ (2016a) suffers from: 1) inconsistencies in categorization, and 2) the incorrect use of confidence intervals to describe quantitative boundaries of protective versus non-protective, forming a dubious scientific foundation for the criteria.
2. We recommend a simplification of categories into either “protective” or “not protective.” As an alternative, we also suggest “least risk” may be substituted for “protective.”
3. The definitions of these categories should be linked, to the extent possible, with quantitative thresholds representing designated use impairment. Quantitative thresholds

identifying "low" versus "high" risk of adverse effects (*e.g.*, HAB alert thresholds) can provide a basis for these quantitative definitions.

4. There are options for taking this analysis into a better statistical and quantitative framework that include a more quantitative conceptual model with functions fit to exceedance curves (*e.g.*, exponential, sigmoidal), the slopes of which can quantitatively provide comparative risks per unit chlorophyll-*a* across effects.
5. Computed confidence intervals require the classification of categories first, so they are not independent estimates of standard errors of the thresholds. Using these estimates of error to determine the confidence of a given threshold is not appropriate. It is reasonable to use a confidence interval or other statistic (*e.g.*, 75th percentile) to more precisely define that risk threshold, but that statistic cannot in and of itself be the basis for the definition, since the variability is often driven by the density of the available data.

Response to Question #5:

1. Arithmetic means of environmental variables are always higher than geometric means.
2. Relative to any threshold associated with an estimated level of exceedance probability, the geometric mean will indicate a lower exceedance probability than the arithmetic mean.
3. The choice of either an arithmetic or geometric mean to compare with the cumulative frequency diagram (CFD) approach will affect the outcome of computed chlorophyll-*a* values/thresholds.
4. In light of the potential for changing assessment methods, statistical correspondence analyses should be established between how risk is currently evaluated and related to chlorophyll-*a*, and under any proposed changes in assessment methods (*e.g.*, the analysis should be performed using both assessment methods). This correspondence analysis will help to ensure that the risk to aquatic life is limited in a manner as originally expected under the current assessment techniques.

Response to Question #6:

1. The desire to replace the current (*i.e.*, CFD) assessment procedure with the proposed approach is logical. While the current approach is scientifically innovative and repeatable, it suffers from: 1) a lack of consistency with the spatial and temporal scales of data aggregation that serves as the basis for decisions on the extent, duration and frequency of the chlorophyll-*a* criteria, 2) the potential for bias based on data density, 3) complexity, and 4) lack of transparency.
2. While there is a clear need for an alternative to or improvement upon the current assessment approach, the proposed alternative approach has major issues that need to be addressed before it can be presumed to be superior:
    a. Although the alternative is simpler and more straightforward, a question remains as to whether it does a better job of assessing designated use attainment, based on quantitative linkages to adverse effects (see Response 2 and 3).

b. The proposed methodology represents an aggressive aggregation of data, with little documentation on the rationale for such aggregation.
c. A six-year assessment window may delay the management response too long to allow efficient trial and pursuit of alternative nutrient management.
d. A more thorough review of the basis for segmentation is suggested, using indicators associated with adverse effects (HABs, low DO, etc.) as well as co-factors known to control ecosystem response to nutrients.
e. The proposal for the alternative approach criticizes the CFD as being prone to high rates of false positive and false negative errors, yet the proposal does not demonstrate that the alternative methods would result in improved performance in these regards.

Response to Question #7:

1. Use of the proposed harmful-effects-based approach in future chlorophyll-*a* criteria work could certainly hold value and has already been demonstrated in studies such as Harding *et al.* (2014).
2. One strength of the harmful-effects approach comes from its localized use of datasets that lead to site-specific segment and seasonally- based analysis of chlorophyll exceedance. In this regard, it would not be wise to take the James River analysis and assume that these exceedance frequency relationships would be directly applicable to other similar salinity segments.
3. Limits on the amounts of available data were found to constrain the utility of the approach for the JRCC. We recommend that appropriate data be collected in future segments of the Chesapeake Bay or other estuaries, in order to best apply this approach to other systems.

Near Term Recommendations:

1. Refine the analysis that supports segmentation of James River, considering tidal excursion distance and including, where possible, response indicators (chlorophyll-*a*, adverse effects such as DO, etc.) and causal factors known to control eutrophication (turbidity, nutrients, etc.). Use this refined segmentation to improve effects-based analyses and the current and/or proposed alternate assessment method.
2. Refine the effects-based analyses by doing the following:
    a. Simplify categories into either "protective" versus "not protective" or "low risk" versus "high risk"
    b. Identify quantitative thresholds for each of the potential pathways of impairment (low DO, HABs, water clarity, etc.) that represent these "low risk" versus "high risk" thresholds. Thresholds can be derived from existing policy (*e.g.*, DO criteria), state, federal or international guidance (*e.g.*, alert versus action levels for HAB toxins), published literature and/or consensus of scientific working groups.

c. Develop conceptual models describing how chlorophyll-*a* links to each pathway of impairment, specifying the important temporal and spatial scales relevant to consider in the analysis (monthly, seasonal, annual, etc.). Use these scales as the basis for aggregating chlorophyll-*a* data on the x-axis of the conditional probability figures.
d. Use continuous statistical models to investigate the relationship between chlorophyll-*a* and each indicator of adverse effects (*e.g.*, HABs, DO). If a significant relationship exists, use that documented statistical relationship to calculate chlorophyll-*a* thresholds associated with "low" and "high" risk. Quantify uncertainty in these values to the extent practicable, using variance analyses. Compare chlorophyll-*a* low and high risk thresholds generated through multiple pathways of impairments, if possible.
e. Explore more flexible models such as quantile regression or hierarchical Bayesian methods to quantify both the probability of exceedance and its standard errors. In particular, although the conditional probability calculation provides an estimate of the risk associated with a chlorophyll-*a* level, it does not assess the confidence of this estimate. The confidence is likely to vary between the metrics due to the spatial heterogeneity of their relation with the chlorophyll-*a*. For example, the DO relation was not homogeneous in space, and the corresponding metric was not very sensitive in the James River segments.

3. Refine assessment approach (either current or proposed Robertson (2016) alternative), by doing the following:
    a. **If the decision is to continue with the current approach**, investigate how well the CFD approach represents spatial and temporal "covariances of attainment." Consider realigning assessment methodology procedures around: i) original conception of how the CFD should be implemented for the James River; ii) new segmentation based on classification analysis (see near-term recommendation #1 above); iii) an investigation of the effect of temporal and spatial scale on the effectiveness of the CFD approach in discovering significant relationships in chlorophyll-a data (this may vary as a function of sampling intensity in a given location/segment); and iv) better congruence between data handling methods and sample size for defining the reference curves and computing the assessment curves.
    b. **If the decision is to pursue an alternative, simpler assessment approach,** consider refining the scientific basis for a proposed approach by doing the following: i) use an alternative segmentation that could result from recommendation #1; ii) investigate how different methods of aggregating chlorophyll-*a* data at a segment scale change the relationship to pathways of adverse effects (HABs events, etc.); iii) use definitions of high- and low-risk adverse effects to justify the rationale for decisions on data aggregation, and iv)

conduct power analyses on existing monitoring data or model output to determine the minimum sample size required for assessment, in consultation with the SAP or some other expert group.

Longer-term Recommendation:

In the long term, we recommend a sampling design consistent with the effects-based approach. The existing monitoring effort should be distributed in a spatially balanced manner (Stevens & Olsen 2004) to improve the predictive capability of the resulting HAB and chlorophyll-*a* data. Experimental design principles such as stratification, or the techniques of compliance points should be consulted to guide the monitoring effort. Data based on sound design principles will enable unbiased estimate of the risk and the uncertainty associated with the current criteria.

1) Clarity of Writing

While considering the two reports that were the focus of this review, the panel found that important context and background information was omitted. Some of this information was available in the large library of supporting documents that accompanied the main documents. We urge the parties following up on this work to develop and provide additional detail and summaries of both policy and management frameworks so that readers can engage with either report without extensive reference to additional documentation. Alternatively stated, the main reports should largely "stand on their own" as they convey the motivation and other background for the document, its main methodologies, results, conclusions and recommendations. Given the volume of the accompanying documentation, which approached 600 pages, it was not feasible for each panelist to read every document. These challenges will become even more consequential during a period of public comment, where clear and effective communication will be even more important.

In the case of VA DEQ (2016a), some details regarding methodology were not included. In particular, the means by which thresholds were selected are not described. Our review response for questions #3 and #4 addresses this concern in detail, but we also encourage the responding parties to ensure that critical methods and steps are detailed in both reports. Our response to question #1 detailed below also underscores the value of insuring that the assumptions and methodology of the conditional probability approach are articulated in an equation framework. This is provided with some detail in the supplemental information, but may be worth considering as a component of the report methods.

The Robertson (2016) report applied some statistical metrics that are not commonly used, such as Calinski-Harabasz pseudo F-statistic. The addition of a few sentences to describe the selection of these tests and a citation supporting similar usage would be helpful.

Finally, there is a full body of research and literature pertaining to selection of criteria and risk assessment. We suggest additional consideration in citing this work. In particular, the Harding *et al.* (2016) paper recently published on Chesapeake Bay chlorophyll criteria is only nominally cited by VA DEQ (2016a), even as it followed a similar approach to exploring harmful effects. We also include additional references in our responses below from other estuarine systems that may benefit from consideration in revision of the reports.

2) Response to Review Questions

For clarity, we have re-ordered our responses to the STAC Review Charge questions (Appendix A) to flow more logically from the most general to the more specific comments and recommendations. We begin with Question 2 and move Question 1 to follow Question 5. Also, because our responses to questions 3 and 4 were similar, we combined those into one comment. Finally, we moved Question 6, which deals primarily with the Robertson (2016) report, to the end of our response, following Question 7.

Question 2: Please comment on the approach's focus on the harmful effects of algae to derive chlorophyll criteria, rather than using reference conditions (as described in Buchanan, 2016) as an additional line of evidence.

Conceptually, the evaluation of chlorophyll-*a* criteria that are protective against the risk of adverse effects is a scientifically sound and reasonable approach, with ample precedent among US EPA and states in setting criteria protective of human and ecological health, including for Chesapeake Bay (Harding *et al.* 2014). The approach is founded on the fundamentals of human health and ecological risk assessment, which has long been used as the basis for public policy (Suter 1993, US EPA 2014). Identifying the risk of adverse outcomes such as toxic blooms provides a clear rationale for regulation that is easy to communicate to managers and to the public, much more so than a reference-based approach. A good example of this is the public awareness and concern surrounding the discovery of microcystin in Toledo's drinking water supplies, following a cyanobacteria bloom in Lake Erie (see: http://www.nytimes.com/2014/08/07/science/cyanobacteria-are-far-from-just-toledos-problem.html). Connecting chlorophyll-*a* criteria to harmful effects anchors these management tools to the restoration outcomes that are desirable to the management community. The empirical nature of the assessment is particularly appropriate for application to specific locations, such as the segment scale selection applied to the James River by VA DEQ (2016a).

Using a reference-based approach also has its merits. It typically draws from a larger spatial scale and challenges the end-user to consider how they will define "reference" conditions, particularly as the baseline shifts with ongoing global climate change. This may help define what is achievable in a restoration context and will also compel discussion and debate regarding restoration trajectories, for which uncertainty and non-linearities are common features (*e.g.*, Duarte *et al.* 2009). Conceptually, it may be useful to consider that the "frequency of threshold exceedance" graphs (*e.g.*, Figure 4 in VA DEQ 2016a) may have "reference" conditions represented at the left-hand side of each graph in the "least risk" category provided the dataset encompasses sufficiently low chlorophyll-*a* conditions. For this reason, it is worthwhile to consider that the reference-based approach is complementary to, rather than an alternative to, an effects-based approach. For both approaches, the empirical relationships are necessarily limited

to the conditions and data available for the analysis. VA DEQ (2016a) states this challenge, but it is worthwhile repeating in this response.

While an effects-based approach has substantial value, the panel recommends refinements to the implementation of this approach documented by VA DEQ (2016a). First, clear definitions of the levels of adverse outcomes (*e.g.*, low DO, HAB toxins) that would be considered "protective" versus "not protective" were not used in the analysis. The rationale for those decisions could be based on existing policy, state, federal, or international guidance, or consensus among scientists on thresholds (*e.g.*, site specific water clarity levels protective of seagrass). Quantitative levels associated with "low risk" and "high risk" of adverse effects can provide a clearer rationale for these categories and related discussions of uncertainty (see Charge Questions #3/4 response). Identification of these risk thresholds can then be used to identify quantitative chlorophyll-*a* values which correspond to those levels of risk, the tradeoffs of which can be clearly communicated to the public. This is superior to identifying whether current criteria are protective or not protective against that risk (VA DEQ 2016a). The outcome using this approach will provide the chlorophyll-*a* level that limits risk ***as intended***, where the current criteria might be minimally protective, with excessive risk of not being protective or alternatively lower than needed. Identifying risk levels associated with low dissolved oxygen and HAB toxins should be fairly straightforward, by using either existing policy or statewide guidance. If no state guidance or adopted policy exists, a review of other state or international guidance (*e.g.*, Sutula and Senn 2016), can be a fair substitute as a basis for discussion with managers. If this literature is deemed insufficient, consultation with an expert panel is also an avenue to support decisions on risk-based thresholds. Harding *et al.* (2014) provides an excellent example of this quantitative approach. There are disconnects throughout the VA DEQ (2016a) report in documenting this link to policy. In addition to recommending a clearer approach for selecting quantitative thresholds for the harmful effect, the selection of the metrics themselves must also be carefully considered. For example, the decision to use algal contribution to suspended particle matter as an indicator of water clarity is an approach that has not been extensively vetted in the peer-reviewed literature.

A second critique is that VA DEQ (2016a) did not clearly articulate the spatial and temporal scales associated with risk of adverse effects and how these were considered in the aggregation of data used in the X- and Y-axis of the conditional probability analysis. For example, toxic HAB blooms may be empirically related to chlorophyll-*a* on monthly timescales, while summertime or annual chlorophyll-*a* may be more strongly associated with low dissolved oxygen (segment annual average versus monthly event average). It is important to justify these decisions made in the effects-based analysis in VA DEQ (2016a) and show to what degree they support decisions on the extent, frequency, and duration specified in the criteria. We encourage the SAP to think through how these spatial and temporal scales, along with their relevant statistics, can be reconciled to provide a more ecologically meaningful, risk-based analysis.

To the panel, these discrepancies spoke to a fundamental missing link in this report between policy-relevant metrics of adverse effects and their corresponding chlorophyll-*a* concentrations. Clearly articulating the relevant spatial and temporal scales in the effects-based analysis will make the appropriate aggregation methodology and other statistical considerations in an assessment methodology easier to understand and define. A lack of clarity in this area, perhaps arising from the complexity of the currently adopted cumulative frequency diagram assessment protocol, makes the validity of the harmful-effects analyses difficult to evaluate. Rather than specify that the effects-based analyses align with the preferred assessment methodology (US EPA 2008), which may be overly burdensome or limiting, the panel recommends looking ahead. As the effects-based analysis is further refined, it is critical to understand that a tractable assessment methodology that limits risk as intended will have to be defined and periodically implemented in perpetuity. At a minimum, it is important to ensure that the way that the chlorophyll-a data are aggregated in analysis such as those described by VA DEQ (2016a) is well understood, especially for data differing in their degree of spatial and temporal resolution, so that it can be taken in account in devising a corresponding protocol for criteria assessment.

Question 3: Please comment on the approach for defining three categories of threshold exceedances as 'protective', 'defensible', and 'not protective' and on the approach for deciding if the categorization of these threshold exceedances are scientifically defensible. Please also comment on the general concept of applying these definitions to make the determination as to whether the existing Virginia chlorophyll-*a* criteria are both protective of the aquatic life designated use and scientifically defensible.

Question 4: "Please comment on the following findings: 'The results of the effects-based analysis suggest that the current criteria are defensible in that they fall below the non-protective range. In most cases, the criteria fall above the upper threshold for low risk indicating that lowering the values of the criteria may result in further improvements in water quality and phytoplankton condition. However in most cases, anticipated reductions in frequency of exceedance at attainment of the low risk threshold were small.'' (p. 36 of VA DEQ *et al*. 2016)

Terminology

VA DEQ (2016a) classified exceedance rates into two categories, "protective" and "not protective" (Figure 4 in VA DEQ 2016a). Subsequently, the "protective category" was subdivided into two additional categories, "least risk" and "defensible." This scheme is subtly different from the three-category scheme implied by charge Question #3. The distinction is important, however, because for the purpose of assessing and classifying waters, there are effectively only two possibilities: either chlorophyll-*a* is too high (*i.e.*, "not protective") or it is not too high (*i.e.*, "protective") in relation to 303(d) listing of impaired waters. Within the category of "protective" it could still be informative to have two categories, as suggested, but it may not be well-advised to call one of them "defensible" since presumably all the classifications

must be defensible (*i.e.*, there should also be a valid scientific rationale for stating that the highest chlorophyll-*a* classification is "not protective"). As an alternative, it may be preferable to call one "protective" and the other "least risk," with the presumption that "least risk" is also protective. The term "defensible" has a specific meaning in a regulatory framework and is inappropriate in this context. Terms such as "least risk", "moderate risk", and "high risk" of threshold exceedance, could be defined in relation to state policy or guidance (*e.g.*, DO criteria, HAB advisory guidance), but these create uncertainty that must be resolved regarding the policy implications of the middle "moderate" category.

Classification Methodology

The approach for classifying chlorophyll-*a* into the three categories is unclear and appears to be fully subjective in VA DEQ (2016a). On page 13, VA DEQ (2016a) noted that "existing criteria were judged to be 'not protective' if falling within the chlorophyll-*a* range where *elevated* threshold exceedance values were observed" [emphasis added]. No operational definition of 'elevated' is provided, however. A major problem for evaluating this question is the conceptual model in Figure 4 (p. 12 in VA DEQ 2016a). The graph appears to have real data but is not quantitative, with arbitrary placement of A and B lines to separate the three zones of risk ("least risk", "defensible", and "not protective"). Line B in Figure 4 is drawn at an apparently arbitrary distance to the left of the two observations that are clearly much higher than the other observations, the placement appears to be arbitrarily placed in a data gap. The three observations at and to the left of Line A are also clearly 'elevated' relative to the three observations for which there were no exceedances. Line A is also an arbitrary distance to the right of the next point to the left, which carries no higher risk than the point intersecting Line A. Whereas Line A is suggested to delineate risk at the "low end of their observed distribution", this distinction is not defined quantitatively and therefore appears to also be arbitrary. Arguably, Figure 4 depicts a linear increase in risk of threshold exceedance with increasing chlorophyll-*a*. On page 18, VA DEQ (2016a) noted that "patterns in the relationships between expected threshold exceedance with mean chlorophyll-*a* were used to infer the limits of lowest risk, defensible and non-protective ranges." However, once again, no method is described. In conclusion, if there is a reproducible method to delineate the range of chlorophyll-*a* values into three categories with respect to risk, it is not clearly presented.

Given the absence of a quantitative approach for classifying chlorophyll-*a*, determining confidence intervals for the thresholds presents a problem. VA DEQ (2016a) notes on page 18, referring to Table 5, that "standard errors associated with these mean values were used to assign confidence intervals to thresholds." This statement refers to standard errors of observations within the categories. However, determining the confidence intervals require the classification of categories first, so they are not independent estimates of standard errors of the thresholds that define the categories. These standard errors may have no valid interpretation with respect to the thresholds and *at best* provide qualitative information. Given observations classified into categories of "least risk" and "non-protective", VA DEQ (2016a) noted that confidence intervals

were computed for group chlorophyll-*a* means, with the suggestion being that the upper and lower confidence intervals for the "least risk" and "non-protective" means are thresholds. This logic is flawed, since the confidence intervals cannot be calculated until after the observations are previously classified.

A suggestion for establishing a quantitative classification methodology is to model the relationship between chlorophyll-*a* and exceedance risk (Figure 1). Such relationships could have a variety of functional forms, as dictated by the data. Given a policy decision regarding the acceptable level of exceedance risk, it should be possible to identify a chlorophyll-*a* level associated with that average risk. Or, using a conditional probability approach (Paul and McDonald 2005, Hollister *et al.* 2008), to calculate the threshold for chlorophyll-*a* above which the risk exceeds the acceptable level. Although we do not provide a specific approach here, it may be possible to derive limits of uncertainty for the threshold (*i.e.*, fiducial limits). To illustrate the selected approach, it is recommended to utilize one or more datasets depicting real and representative data (*e.g.*, frequency of *Cochlodinium* threshold vs chlorophyll-*a*) where correcting this lack of quantitative classification might include some model fitting. Many of the exceedance effects appear to have exponential or sigmoidal relationships to chlorophyll-*a*. In the case of the apparent data in VA DEQ (2016a) Figure 4, fitting an increasing exponential function of exceedance risk could provide estimates of the slope (increasing risk of exceedance per unit chlorophyll-*a*) which could be used as quantitative thresholds across effects. Other effects may be better described by sigmoidal or decreasing exponential functions. We suggest that a real and representative example should be used in the place of VA DEQ (2016a) Figure 4. Our request is for one representative graphic example for each type of effect (positive or negative exponential, sigmoidal, etc.), supplemented with a table giving parameter values for all effects.

A further issue noted by VA DEQ (2016a) is that very narrow confidence intervals are associated with abundant data, resulting in narrower confidence intervals that impact the thresholds. Using continuous underway sampling (*e.g.*, the "Dataflow" approach cited in the report) resulted in narrower confidence bands. We posit that computing standard errors or standard deviations from measurements obtained through continuous underway sampling as if they are independent is not defensible given the strong spatial dependence these data will have. Using a measure of variability (*i.e.*, standard deviation) rather than precision (*i.e.*, standard error) could address this problem if the dataset was composed of independent data, although variability within groups does not characterize uncertainty in the thresholds dividing the groups. Beginning the analyses with the objective of estimating "protective" and "non-protective" categories should be a first step in the conditional probability approach.

In regard to determining whether these definitions can help to determine if the existing Virginia chlorophyll-*a* criteria are protective, we emphasize that without quantitative methods for separating the various categories it is difficult to scientifically defend the placement of a line on the exceedance graphs. We repeat this here as a summary of our discussion above, which is

focused on the amount of reduced risk associated with lower chlorophyll-*a* criteria, and to specifically address Review Question #4. One issue with answering this question, as noted previously, is that the panel feels that "protective" versus "not protective" has not been properly defined, based on consultation with managers on acceptable levels of risk associated with low DO, HAB toxins, water clarity, etc. (see answers to Comments 2 and above). Without a quantitative conceptual model of how this will be evaluated, it is difficult to answer this question. If Figure 4 of VA DEQ (2016a) can be reasonably improved quantitatively, then any new criteria should be at least as protective as the current criteria. Due to weaknesses in the underlying analyses of VA DEQ (2016a), it is difficult to determine whether lowering the chlorophyll-*a* criteria would significantly reduce the risk of exceedances. The panel recommends a refinement of the analyses to address these issues, in order to clarify these risk relationships and their interpretation as the basis for chlorophyll-*a* criteria decision making. This would need to be rectified before the review panel felt the criteria could be critically evaluated. Such a critical evaluation could then focus on presentation of a fit for a quantitative model and associated uncertainty.

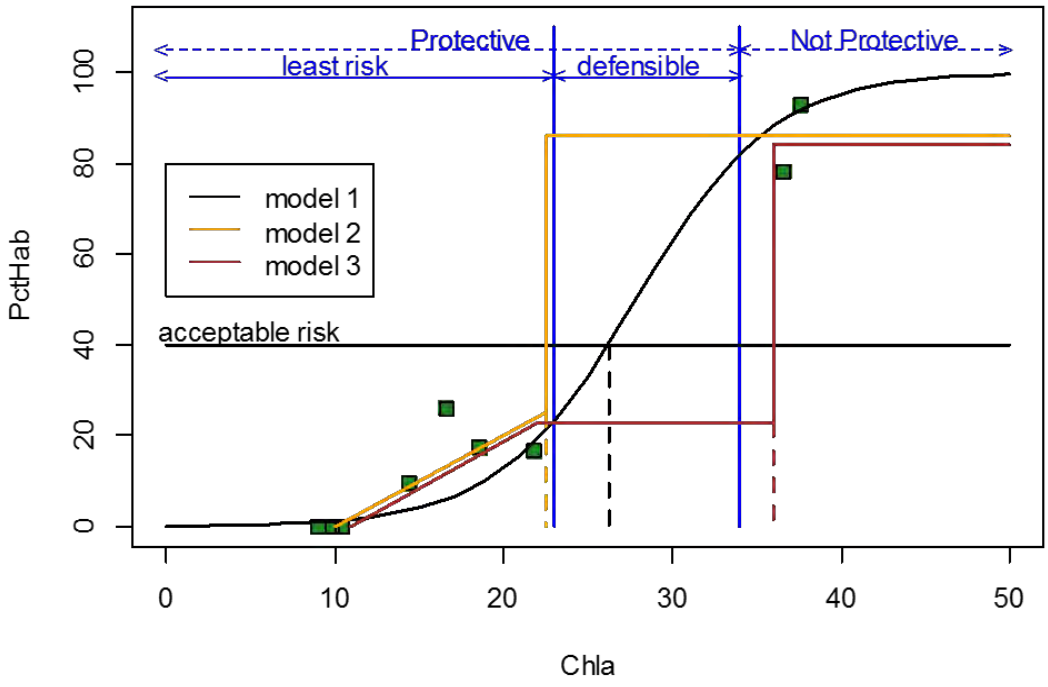

Figure 1. Empirical relationships between percent HAB exceedance and average chlorophyll-*a* presenting our suggested classification of protective and not protective against the VA DEQ (2016a) classifications. Multiple models are represented and can be set against a pre-defined acceptable risk.

Protectiveness of Existing Criteria

Since the levels of acceptable risk do not appear to be clearly defined, statements regarding the protectiveness of existing criteria rest on a shaky foundation. Neglecting this fact, the panel nonetheless encourages the approach of relating the existing criteria to effects metrics such as

HABs as a useful step forward, as has been noted by others (Comments by Clifton Bell, April 18, 2016). Given levels of acceptable exceedance risk, it may be possible to define thresholds and subsequently evaluate observations meeting the current criteria that would be expected to be associated with acceptable exceedance risk for the endpoints. Doing so would establish if the existing criteria are protective, but would not indicate if a higher threshold would also be protective.

In contrast, statements to the effect that there are significant differences in HAB endpoints among subsets of chlorophyll-*a* above and below the existing chlorophyll-*a* criteria (*e.g.*, Page 17 VA DEQ 2016a) provide no information regarding the appropriateness of the threshold. Rather these statistical differences simply indicate that there is a relationship between chlorophyll-*a* and the HAB endpoint. Group mean differences (*i.e.*, mean exceedance rate when chlorophyll-*a* is above or below the threshold) would emerge in that case given any arbitrary threshold delineating the groups.

Methods such as conditional probability analysis or quantile regression can be used to estimate the mean probability and confidence intervals of reaching specific "protective" versus "at risk" thresholds. An improved approach to this risk-based analysis of adverse effects would be to define what threshold and corresponding statistic (*e.g.*, mean, 95th percentile) would be used to define "protective" versus "non-protective." Policymakers should be briefed on the associated uncertainty and their consensus, to the extent possible, incorporated into *a priori* decisions on interpretation of analyses. Examples of these metrics can be found for conditional probability analysis of HAB risk and quantile regression of low DO on increasing chlorophyll-*a* for San Francisco Bay (Sutula and Senn 2016, Sutula *et al.* in review).

Question 1: Please comment on the scientific basis for applying a combined probability approach to derive expected frequencies of threshold exceedance as a function of mean chlorophyll-*a* to determine whether attainment of these criteria would result in low rates of threshold exceedance.

The James River Science Advisory Panel (SAP) proposed an effect-based approach to consider multiple metrics that link the chlorophyll-*a* concentration with indices of aquatic life designated uses as reported in VA DEQ (2016a). The approach utilized a decade of data on water quality, phytoplankton community, and occurrence of harmful algae. For each metric, the expected frequencies of threshold exceedance were computed as a function of chlorophyll-*a* using a conditional probability approach. The sampling efforts are intensive for chlorophyll-*a* and corresponding metrics, which provides rich data and sound bases for comprehensive effects-based analyses. The conditional probability approach is non-parametric and based on mild assumptions regarding the monotonic functional dependence between chlorophyll-*a* concentration and target metrics. We feel there are benefits to this approach and detail our

examination of the combined probability approach and its assumption in this response. There are, however, spatial misalignments between monitoring data sets, which could introduce bias in the expected frequency calculation. There is also a need to derive model-based estimates of standard error. We also make some recommendation regarding cost-effective monitoring for effects-based analyses. Details of this summarized response to question #1 are provided below. In summary, the conditional probability approach implicitly assumed spatial homogeneity of the relations between chlorophyll-*a* and various end points, and pulled together monitoring data of diverse spatiotemporal resolution and extent (Table 1).

Conditional Probability Approach

We feel that understanding the scientific basis for the conditional probability approach benefits from articulating the assumptions of the VA DEQ (2016a) analyses using equations, a feature that was not included in the methods description of the report, but provided in one example calculation in the supplemental information. We expand upon this below, articulating our understanding of underlying assumptions as an important step in assessing the validity of applying this approach to the James River.

Supplemental Materials and the Conditional Probability Approach

In this context, let TE_{year} denote a binary random variable (RV) of threshold exceedance. This RV is defined for each year given a specific combination of segment and season. Thus, year is the assessment unit. Let CHL_{year} denote a categorical RV of chlorophyll-*a* (CHLa) distribution from the same study domain, where x denotes the bin of chlorophyll-*a* values in 10 µg/l intervals $x \in \{(0,10], (10,20], (20,30], \ldots\}$ and B denotes the total number of bins. The conditional probability is then generally defined following this equation:

$$\Pr(TE_{year}) = \sum_{x=1}^{B} \Pr(TE_{year} | CHL_{year} = x) \times \Pr(CHL_{year} = x). \qquad (Eq.\ 1)$$

Both probability statements on the right-hand-side of (1) are specific to the entire segment and water column spatially.

To facilitate estimation of the probability, three assumptions were made in the report: (1) the spatial domain was restricted to the surface water (*e.g., let* Pr_s *denote the probability specific to surface water (<1.5 m), hence* $Pr(CHL_{year} = x) = Pr_s(CHL_{year} = x)$ *in Equation (1));* (2) the conditional probability $\Pr(TE_{year} | CHL_{year} = x)$ does not vary from year to year, hence we can combine data across years; (3) the sampled sites are assumed to be representative of the entire surface layer within the segment in terms of the conditional probability $\Pr(TE|CHL=x)$. The last assumption addresses the spatial mis-alignment between sampled sites and the entire shallow segment. With these assumptions, the conditional probability $\Pr(TE|CHL=x)$ was estimated at the sampled sites, and extrapolated to the entire segment by multiplying by the marginal probability $\Pr_s(CHL_{year} = x)$.

Spatial Misalignment

The validity of these assumptions depends on the sampling design. For non-probabilistic sampling designs, bias can be introduced by extrapolating the relationship observed at sampled sites spatially to the entire segment. To explore whether bias was an issue in the application of this approach to the James River (VA DEQ 2016a), we compared the sampling characteristics of the data used to derived the conditional probability (termed as source) and those used to derive marginal probability (termed as target) to evaluate the degree of uncertainty in the extrapolation.

The spatial data and temporal resolution of the datasets were extracted from the report (VA DEQ 2016a) for selected metrics. The levels of spatiotemporal misalignments vary with the metrics (Table 1). For harmful algae bloom (*Cochlodinium*) in the lower James, the spatiotemporal locations and timestamps for the source and target are similar. For water quality conditions such as DO and pH, however, the misalignment is more obvious. The conditional probability was derived by averaging over the temporal variance component captured by continuous monitoring ("ConMon") sampling methods, while the marginal distribution was developed by averaging over the spatiotemporal variance component measured by the continuous underway sampling approach. Conceptually, these two variance components are different and can affect the expected frequency calculation. Furthermore, because the system is tidal, the temporal and spatial variance of the sampling measurements become mixed as water moves back and forth past a fixed continuous monitoring location, and the same mixing of variance components occurs due to non-synopticity of continuous underway measurements

Validation Using a Hydrodynamic-Biogeochemical Model

Validation of the assumptions applied in VA DEQ (2016a) would require complete sampling of the entire system, which is unrealistic in practice but possible in a virtual sense with numerical simulations. As part of this review, we carried out this simulation approach in an attempt to test these assumptions. Thus, DO and chlorophyll-*a* simulations were conducted on a 80 × 120 grid cell over the entire Chesapeake Bay water shed using output from an implementation of the ROMS-RCA hydrodynamic-biogeochemical model (Testa *et al.* 2014). The simulations from James River segments were extracted every 3 hours over the entire water column. Values corresponding to the ConMon stations were identified to calculate the conditional probability. As a reference, we also calculated the conditional probability using all surface cells within the segment. The calculations were conducted for model output from Spring and Summer of 2004 and 2005 and corresponding empirical measurements from the ConMon dataset.

There are subtle differences in the conditional probability distributions for the mesohaline segment in summer (Tables 2 and 3). The exceedance frequency for the entire surface layer tends to be higher than for cells corresponding to the ConMon stations. In fact, no exceedance was observed in surface cells containing the ConMon stations. If this difference is real, using the ConMon based conditional distribution to derive threshold exceedance probability could lead to an underestimate of the risk. The conditional distributions also vary from year to year. The exceedance probability in bins (10-20], (20,30] and (30,40] were larger in 2004 than 2005. Assuming validity of the numerical model, ignoring the year to year variability in the conditional probability distribution could also introduce extra variability in the expected frequency calculation.

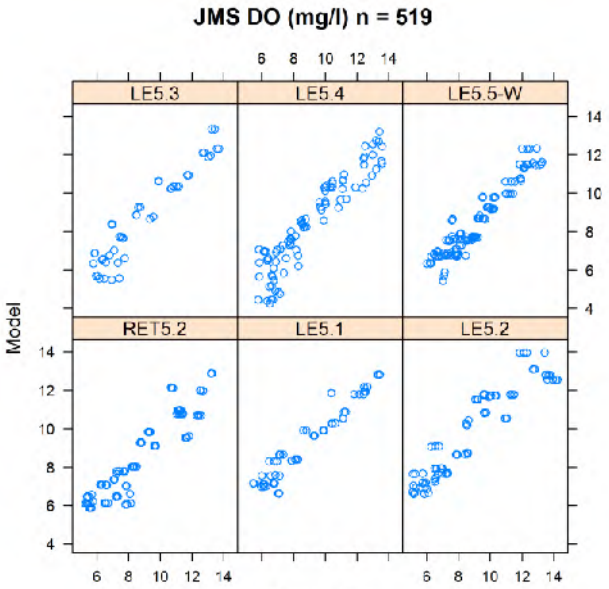

Figure 2. CBP monthly DO data from stations in the James River, and the closest ROMS-RCA model

The performance of the ROMS-RCA model is unknown for the James River. Thus these numbers should be interpreted in the light of the underlying model uncertainty. Skill assessment of the model suggests that the DO simulations are essentially un-biased for oligohaline, mesohaline and polyhaline segments of the James (Figure 2). *Thus the spatially homogeneous relation between DO and chlorophyll-a underlying the lower James assessment* (VA DEQ 2016a) *is not verified by numerical simulation*. We see that the chlorophyll-*a* simulations are biased (Figure 3). Therefore, we used empirical data from ConMon and CBP fixed stations to compare the conditional distribution of chlorophyll-*a* and water quality conditions (DO, pH, temperature). We understand that the chlorophyll-*a* data collected using sondes is based on fluorescence measurements, versus direct measurement at the fixed station.

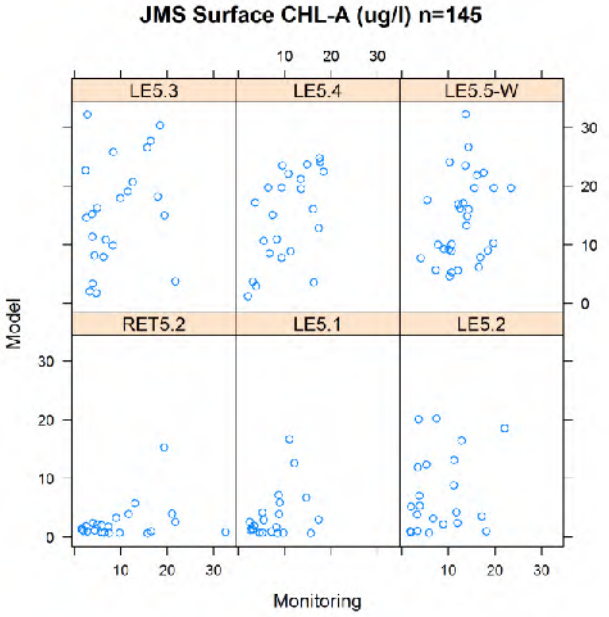

Figure 3. CBP monthly chlorophyll-a data at the surface from fixed stations in the James River, and the closest ROMS-RCA model output.

Validation Using Empirical Data

The ConMon station data used in the VA DEQ (2016a) analyses were temporally linked to the nearest fixed station and the long term measurements at the surface. The scatter plots of chlorophyll-*a* and other variables were used to compare the data from ConMon and nearest fixed

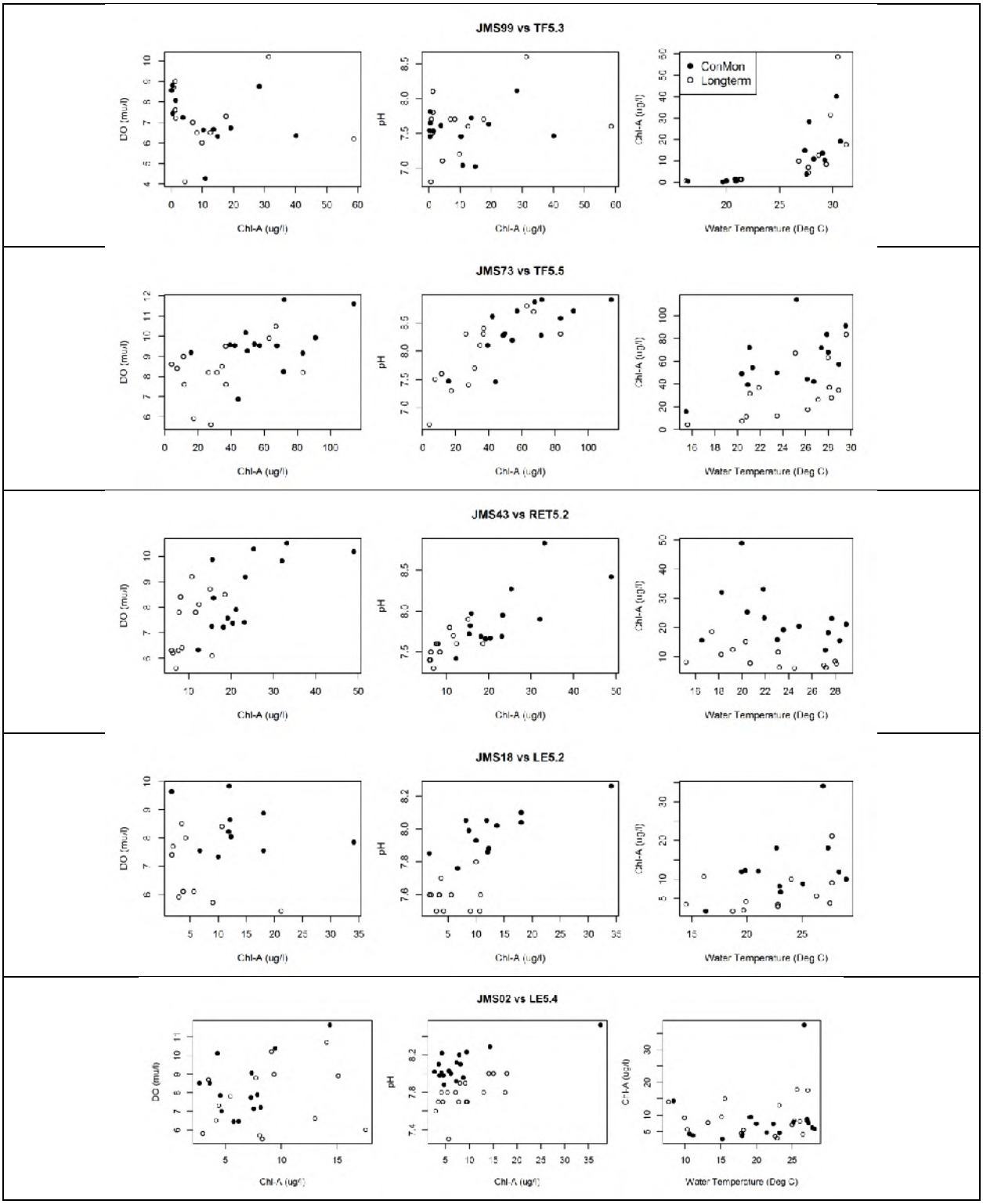

Figure 4. Chlorophyll-*a*, dissolved oxygen, pH and water temperature from the CBP monthly long term monitoring station (2006 – 2008) in the upper tidal fresh, lower tidal fresh, oligohaline, mesohaline and polyhaline segment of the James River, and the collocated (within 1 day) variables (averaged to daily resolution) from the nearest continuous monitoring station.

stations. The scatter plots are spatially different (Figure 4). The ConMon and fixed station data are more similar in the upper tidal fresh segment, while in the lower James, the scatters were quite different. The fact that chlorophyll-*a* appears to be higher at lower temperatures in the ConMon stations (column 3; Figure 4) is consistent with our understanding that shallow waters tend to have higher chlorophyll-*a* than the deeper waters - at least in the near-surface. Due to sample size limitation, comparison was not further stratified by season. The differences in these relationships as a function of sampling location speaks to the importance of aligning the datasets before proceeding with an empirical analysis seeking to address the threshold exceedance issues outlined by VA DEQ (2016a).

Suggested Use of a New Model

The binning of chlorophyll-*a* data provides a non-parametric way of estimating threshold exceedance. The underlying assumption of the association between aligned chlorophyll-*a* and the harmful-effects metric is quite general. The condition of a significant *t*-test result (VA DEQ 2016) given the current chlorophyll-*a* threshold implies a monotonic relation between the metric and chlorophyll-*a*. However, the cost of such generality is the lack of standard error estimates for the expected frequency of threshold exceedance. We strongly emphasize that these standard error estimates are essential for estimating the false positive and false negative rates. A monotonic relationship between chlorophyll-*a* and an associated metric must be a condition to taking the next step of performing the threshold exceedance analysis, because it is possible that available data are insufficient to characterize the relationship. We suggest that a model should be incorporated for understanding the cost-effectiveness and protectiveness of the assessment process.

VA DEQ (2016a) explored logit and *lowess* models, but commented on the apparent lack of fit to the data, which could indicate a poor choice of model. This could be an example where data are not sufficient to evaluate the relationship between chlorophyll-*a* and harmful effects. Binning data by chlorophyll-*a* categories could be hiding a non-monotonic response, which should be modeled through a non-linear approach such as Generalized Additive Models (GAM), or it could indicate lack of sufficient data. The lack of fit is not an argument against using of models, but only an indication of the variability within the data. It does suggest poor predictive capability of the models, but the inferential goal should be accurate estimation of the conditional and marginal probabilities in Equation (1). There is evidence from geostatistics literature that a model with poor predictive performance could generate unbiased estimation with minimal variance (Brus and DeGruijter 1993). As pointed out by VA DEQ (2016a), the use of a model also enables binning of data when data are relatively sparse in some range of the chlorophyll-*a* values. Thus, the use of model could still be useful to assess the uncertainties of the threshold exceedance estimates.

A hierarchical Bayesian model can be employed to quantify the diel, spatial, and weekly variance components of chlorophyll-*a* (CHLa) concentration based on the existing data from

ConMon, DataFlow and long term monitoring efforts. Such a model might resemble the following:

$$\text{latent} = \text{diel} + \text{spatial} + \text{weekly}$$

$$\log(\text{CHLa}) = \text{latent} + \text{noise}$$

$$\log(\text{HAB}) = \text{confounder} + f(\text{latent}) + \text{error}.$$

The hierarchical model specifies a latent random process to model the chlorophyll-*a* at multiple scales. The latent field is linked to empirical HAB data through a semi-parametric additive model with chlorophyll-*a* and confounders such as light availability. This model addresses the misalignments between HAB and chlorophyll-*a* through borrowing information at multiple scales. Statistical inference should be conducted in the Bayesian framework due to the high dimensionality of the data sets. We emphasize that Bayesian modeling as above still relies on similar assumptions as the conditional probability approach in Eq. (1). It does not fully address the limitations in the data collection. Thus model diagnostics should be conducted before extrapolating the model. If the diagnostics do not suggest lack-of-fit, however, the resulting standard error estimates will have more desirable properties and could be useful for the JRCC.

Suggested Cost Effective Future Monitoring

Extensive monitoring efforts have generated a wealth of information on chlorophyll-*a* in the James River. VA DEQ (2016a) recognizes and incorporated this strength in their work on the JRCC. We recommend further synthesis of the data from ConMon, DataFlow and fixed stations to quantify the variance components of chlorophyll-*a* at temporal (diel, weekly) and spatial scales. These variance estimates would enable a power analysis to quantify the existing capabilities of estimating chlorophyll-*a* concentration at the spatiotemporal scales defined in the criteria assessment.

In addition to data synthesis, VA DEQ (2016a) also describes monitoring of HAB. We recommend such efforts be continued, and if possible replicated in other segments of the Bay. Spatial sampling design could be considered to conduct cost effective monitoring and generate data consistent with the conditional probability approach. A fixed station design was used for HAB monitoring, with stratification *between* segments. Five stations were aligned with the long term stations; the other seven stations were not documented clearly in VA DEQ *et al.* (2016). We recommend further stratification *within* segment based on auxiliary information. One potential useful approach for effect-based analysis is to incorporate DataFlow in the design stage. Specifically, the surface chlorophyll-*a* could be interpolated over the lower James segments to identify water bodies with low, median and high chlorophyll-*a*, as well as their variances. The sites can then be allocated in terms of the estimated chlorophyll-*a* distribution, and proportional to the variances observed in a spatial balanced manner. Even though the stratification variable is subject to uncertainty due to the flashy nature of chlorophyll-*a*, such stratification may still

improve estimation of the conditional probability between HAB and chlorophyll-*a* within a segment, and in term lead to unbiased estimates the expected exceedance frequency.

Table 1. Metrics and data resolutions used to develop empirical relationships linking chlorophyll-*a* with threats to designated uses. *Source* denotes the data used for conditional probability of exceedance given chlorophyll-*a*, *Target* denotes the data used for marginal probability calculation for chlorophyll-*a*.

		Spatial Support		Temporal Resolution	
Metric	Segments	Source	Target	Source	Target
DO, pH	Tidal Fresh	One ConMon station per segment	12 stations (inc. 5 CBP stations)	Continuous 06-08	Weekly during spring and summer 09-14
Clarity, PIBI		CBP stations		Monthly 85-14	
HAB *(Microystin)*		JMS85,JMS99 in UTF, JMS75,JMS69,JMS56 & Rice in LTF		Weekly 11-14	
DO, pH	Lower James	One ConMon station per segment	Continuous	Continuous 06-08	Every 1-2 weeks during spring (March-May) and summer (July-September) 09-14
Clarity, PIBI		CBP stations		Monthly 85-14	
HAB *(Cochlrodinium)*		Continuous		Every 1-2 weeks during spring (March-May) and summer (July-September) 09-14	

Table 2. Conditional probability of threshold exceedance (DO average <5 mg/l) in mesohaline James segment during Summer (July-September) 2004: (a) using all Surface layer with water depth less than 2 m; (b) using only cell with adjacent ConMon stations with 1 km.

CHL Bin µg/L	(a) Surface		(b) ConMon	
	N	%	N	%
0-10	24	0.0%	-	0.0%
10 - 20	43	0.7%	-	0.0%
20 - 30	4	0.9%	-	0.0%
30 - 40	6	3.5%	-	0.0%
40 - 50	-	0.0%	-	0.0%
50-60	-	0.0%	-	0.0%
60-70	-	0.0%	-	0.0%
70-80	-	0.0%	-	0.0%
80-90	-	0.0%	-	0.0%
90-100	-	0.0%	-	0.0%
>100	-	0.0%	-	0.0%

Table 3. Conditional probability of threshold exceedance (DO average <5 mg/l) in mesohaline James segment during Summer (July-September) 2005: (a) using all Surface layer with water depth less than 2 m; (b) using only cell with adjacent ConMon stations with 1 km.

CHL Bin µg/L	(a) Surface		(b) ConMon	
	N	%	N	%
0-10	81	0.1%	-	0.0%
10 - 20	13	0.2%	-	0.0%
20 - 30	-	0.0%	-	0.0%
30 - 40	-	0.0%	-	0.0%
40 - 50	2	1.8%	-	0.0%
50-60	-	0.0%	-	0.0%
60-70	-	0.0%	-	0.0%
70-80	-	0.0%	-	0.0%
80-90	-	0.0%	-	0.0%
90-100	-	0.0%	-	0.0%
>100	-	0.0%	-	0.0%

Question 5: Please comment on the finding that "the criteria were found to be less protective when interpreted as geometric means, indicating that conclusions regarding protectiveness are somewhat sensitive to the methodology by which attainment of the criteria is determined.

Arithmetic vs. Geometric Means

Arithmetic means of environmental variables are always higher than geometric means. Thus, given any set of observed chlorophyll-*a*, the geometric mean will always be lower than an arithmetic mean. Relative to any threshold associated with an estimated level of exceedance probability, the geometric mean will indicate a lower exceedance probability than the arithmetic mean. In this regard, what is probably most important is that the exceedance probability associated with the threshold is neither higher nor lower than expected. This is most likely accomplished by using similar statistics for developing relationships and applying them.

One example in which arithmetic means are often preferred is in evaluation of loadings from rivers. This is because the product of arithmetic mean discharge and concentration estimates the total loading, whereas the geometric mean is less unless concentration and flow are invariant. Since HAB abundance is not a direct response to chlorophyll-*a* (HAB species are often not the dominant algal species), HABs risk may be associated with cumulative nutrient effect, which (as noted in the supplementary information VA DEQ 2016a) is better reflected in arithmetic means, "even for log-normally distributed variables."

Effect of Cumulative Frequency Approach on the Question

The existing methodology of applying the cumulative frequency diagram approach involves classifying season-year arithmetic means as meeting or not meeting a threshold on a cell-by-cell basis in the assessment layer, an interpolated field of chlorophyll-*a*. The percentage of cells in the interpolation violating the criteria in the segment is then calculated and used to determine the cumulative probability of the space violation rate. Thus, the CFD assessment approach combines an arithmetic mean (per cell) with a non-parametric approach (per segment) because the number of cells having annual arithmetic means is counted and it does not matter whether a cell exceeded by a large or small margin. On the other hand, creating the assessment layer by interpolation could generate a larger areal extent of exceedance if one of the included observations is very large, depending on how the interpolation is accomplished. Overall, it is not entirely possible to evaluate which averaging method would be most similar to the CFD result should a traditional, non-CFD assessment methodology be applied.

To ensure that the risk to aquatic life is limited as expected, a statistical correspondence should be established between how risk is evaluated and related to chlorophyll-*a* in criteria development and future assessment of chlorophyll-*a*. It may be possible to utilize the along-track samples, continuous monitoring data and traditional monitoring data together to establish a locally-calibrated relationship between spatial-temporal measures of attainment expressed via the CFD and attainment measures that can be evaluated more easily, such as a time series at specific points of observation, provided those points are selected in advance. Such points have been

called “compliance points.” A compliance point approach could be considered to combine biologically meaningful measures of HAB risk in space and time and the need to assess criteria attainment in a consistent and replicable way at reasonable cost.

Question 7: Please comment on whether the scientific basis and procedures described within the Scientific Advisory Panel’s report could be used to derive new chlorophyll-*a* criteria for application to other tidal habitats within Chesapeake Bay with the same salinity regimes and provide similar levels of protection of aquatic life.

To clarify this comment, we emphasize that it would not be appropriate to take the same distribution of data and resulting relationships from the James River analysis and apply the resulting protective or non-protective thresholds to other locations. We feel the strength of the harmful-effects analyses comes from application to localized, segment based spatial scales. The empirical relationships available using the James River reflect the characteristic relationship between chlorophyll-*a* and a given harmful effect variable, and will be specific to the James River estuary and its particular physiographic conditions. The tidal fresh Potomac and tidal fresh James are likely to share some similarities, but it is unlikely that the same frequency distribution would be appropriate for both systems comparing chlorophyll-*a* to dissolved oxygen, for example.

However, if the request for comment is in regard to whether such a general approach might be suitable in other segments and locations, we agree that determining whether existing chlorophyll-*a* criteria in other portions of the Bay are protective against harmful effects associated with HABs, dissolved oxygen, water clarity, or pH conditions would provide compelling analyses. A first step, as was attempted in this study, is to evaluate whether current criteria are protective. Examining the frequency distributions of various metrics of harmful effects in relationship to chlorophyll-*a* can then illustrate whether a) a relationship exists, and b) whether clear thresholds emerge from the data that might help to inform policy decisions regarding a chlorophyll-*a* criteria.

There is precedence for this approach. Harding *et al*. (2014) present an extensive effort in this regard. VA DEQ (2016a) reference this peer-reviewed paper once, although it is only referred to in terms of the difficulties of achieving such relationships. However, Harding *et al.* (2014) present a number of Bay-wide examples where they successfully derived chlorophyll criteria in relationship to light attenuation, dissolved oxygen, and HABs. In this paper, reference conditions derived from measurements in the Chesapeake Bay from the 1960s and 1970s were examined against harmful effects in these three categories – combining the two approaches in a complementary way in line with our thinking described in response to Question 2. Certainly, data such as those from the main stem of the Chesapeake Bay evaluated by Harding *et al.* (2014)

and pictured in Figure 5 for bottom dissolved oxygen suggest that there are empirical relationships that can be leveraged across the Bay.

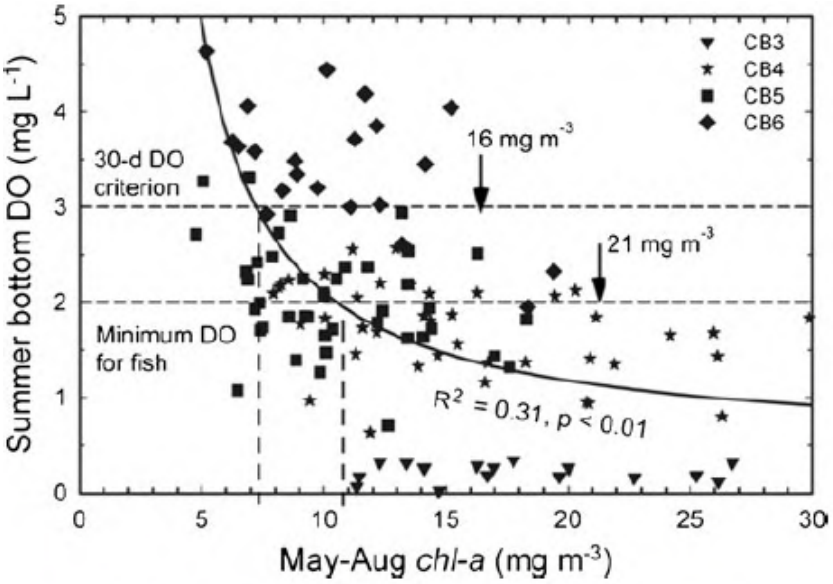

Fig. 5 Non-linear regression (negative, hyperbolic) of mean bottom-layer DO on mean May–Aug surface *chl-a* for four stations in the oligohaline (CB3), mesohaline (CB4, CB5), and polyhaline (CB6) salinity zones of the bay. *Vertical arrows* show surface *chl-a* corresponding to minimum DO criteria for fish (2 mg L^{-1}) and the 30-day DO criterion (3 mg L^{-1}). The 30-day mean DO criterion is based on deep-water DO concentrations that protect against losses in egg survival and larval recruitment of fish species inhabiting subpycnocline waters during summer months (US Environmental Protection Agency 2003). *Dashed vertical lines* show surface *chl-a* corresponding to intersections of the regression line with these DO criteria. The R^2 value and significance level of the regression are shown on the panel

Figure 5. Reproduction of Figure 5 from Harding et al. (2014)

It is likely that applying this approach on a segment scale will be limited only by the availability of datasets. In particular, computing the Phytoplankton Index of Biotic Integrity (PIBI) and other phytoplankton community metrics may be challenging because of the lower density of phytoplankton monitoring stations. We also caution that application of this approach would be recommended only if the challenges we identify in other portions of this report are addressed. In particular, considering alignment of spatial data, quantitative approaches for demarking thresholds, and a necessary feedback between policy-makers and the scientists performing the analysis to determine what level of harmful effects are considered tolerable so that determination of thresholds will fit into desired outcomes.

Whether such an application of separate analyses in other regions of the Chesapeake Bay will result in "similar levels of protection for aquatic life" depends on the distribution of data from those locations and the associated variability between relevant variables. It is possible, for example, that relating chlorophyll-*a* to harmful algal blooms in other segments will not result in meaningful relationships, in which case the probability that a "protective" threshold can be used reliably will be very low.

Question 6: Please comment on the scientific basis for replacing the current chlorophyll-*a* criteria attainment assessment procedures with the proposed alternative chlorophyll-a criteria attainment assessment procedures.

In order to comment on the scientific basis for replacing the current assessment procedure (hereto referred as the CFD approach) with the proposed alternative approach, the panel agreed on a set of evaluation criteria by which to judge both the CFD and the proposed alternative approach:

- Transparent, repeatable, with minimal potential for false negatives or false positives.
- Defines how to assess magnitude, extent, duration, and frequency
- Demonstrable linkage back to policy decision, with consistent use of spatial and temporal scales that connect back to primary analysis that served as the basis for the policy decision.

In the sections below, we first review the scientific basis for the current CFD approach (Secor *et al.* 2006, CBP 2008, VA DEQ 2016b) and then attend to the alternative assessment approach (Robertson 2016).

Evaluation of Current Assessment Procedures (a.k.a. CFD Approach)

The Panel recognizes the current CFD assessment procedure to be scientifically innovative and repeatable (Secor *et al.* 2006, CBP 2008, VA DEQ 2016b). More conventional assessment procedures usually focus solely on temporal exceedance frequency, since spatial uniformity of attainment in an assessment unit is typically averaged or assumed or the density of data required is not available. However, the current assessment methodology suffers from two major (perceived) weaknesses: 1) lack of consistent use of spatial and temporal data aggregation that link back to primary analyses that serves as the basis of the extent, duration, and frequency of the chlorophyll-*a* criteria; and 2) lack of transparency and potential for bias.

The JRCC is defined as a seasonal mean, based on primary analyses that varied – depending on the lines of evidence – from annual means, seasonal means, monthly values and the 90th percentile as reproduced in Figure 6 (Harding *et al.* 2014). It is not clear, however, based on initial review of supporting materials, what the rationale is for the choice of a seasonal mean value for the JRCC. This rationale would be important, because it speaks to the underlying pathways of impairment and the assessment methodology should have a consistency with temporal and spatial scales linked to that impairment. This is not immediately obvious in the documentation provided.

That issue notwithstanding, the current assessment methodology requires a seasonal geometric mean of interpolated values from each grid cell, then a temporal exceedance frequency calculated using the spatial exceedance frequency. It seems odd that the assessment is done seasonally by grid cell. The concept of a grid cell as the unit of assessment seems appropriate

for benthic invertebrates, but not at all appropriate for pelagic parameters like chlorophyll-*a*, since physical and biogeochemical processes can create highly spatially variable conditions that cofound spatial homogeneity and integrity of a grid cell as a sampling unit. Buchanan (2014) notes that exceedance frequencies are not strongly controlled by salinity or locational features, but rather the idea that chlorophyll-*a* concentrations and bloom frequencies are most strongly controlled by the water quality conditions surrounding the phytoplankton population. Given this, it would make more sense to interpolate data for each cruise to estimate a snapshot of the chlorophyll response surface for that date, then determine percent attainment for that event. Percent attainment would then be ranked and assigned quantiles for plotting on the percent of time by percent of space plane. This previous approach would provide more points on each assessment curve, reducing an issue identified and discussed below. We understand that such an approach was originally conceived for the JRCC. Previously STAC recommended increased

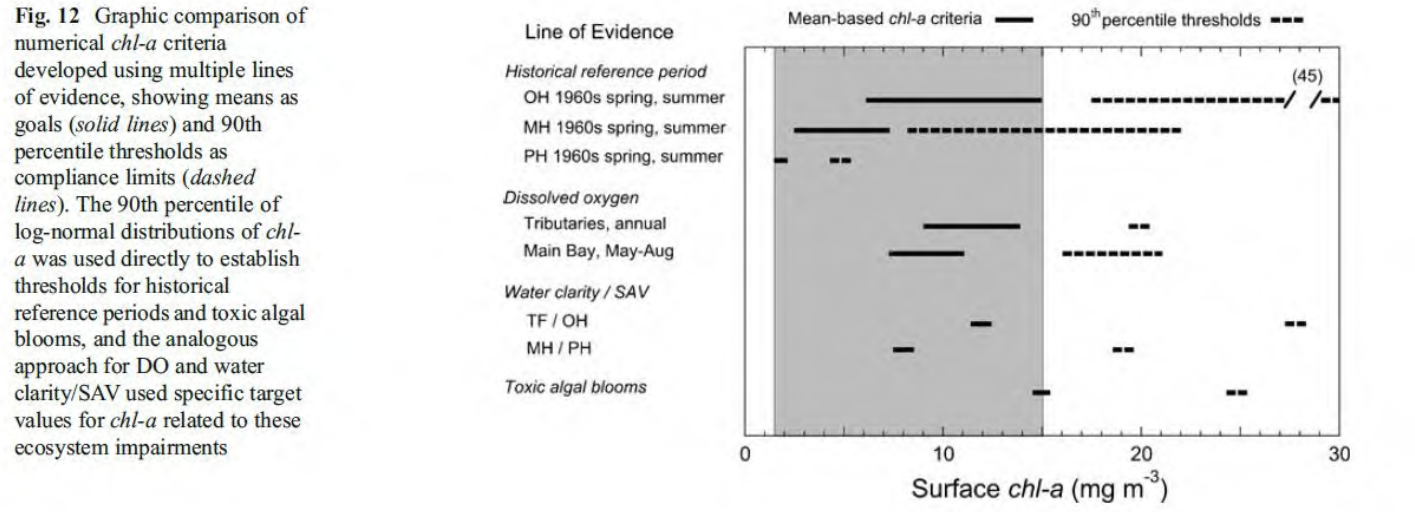

Fig. 12 Graphic comparison of numerical *chl-a* criteria developed using multiple lines of evidence, showing means as goals (*solid lines*) and 90th percentile thresholds as compliance limits (*dashed lines*). The 90th percentile of log-normal distributions of *chl-a* was used directly to establish thresholds for historical reference periods and toxic algal blooms, and the analogous approach for DO and water clarity/SAV used specific target values for *chl-a* related to these ecosystem impairments

Figure 6. Reproduction of Harding et al. (2014) Figure 12.

scrutiny to understand how well the CFD approach represents spatial and temporal "covariances of attainment" (Secor *et al.* 2006). This is important and should be a strong recommendation that supports decision making on any new assessment procedures. The SAP report that is the subject of much of the rest of this panel's review (VA DEQa 2016) features linkage analysis to those lines of evidence (HABs, DO, water clarity, etc.) that indicate use protection and impairment, but it appears neither assessment methodology was used to generate chlorophyll-*a* values used on the x-axes of those analyses. This panel understands that the critique of the current approach and proposal for a new approach to assess attainment with the JRCC (Robertson 2016, VA DEQb) were not linked with the SAP study (VA DEQ 2016a). Thus this improved linkage analyses represents a recommendation for future consideration.

The current assessment method suffers from a lack of transparency. It is computationally intensive and complicated, with the potential for bias introduced: 1) by the number of stations used in the interpolation of monitoring data, especially using limited numbers of fixed stations,

and in 2) the approach and supporting data used to construct the biological reference curve. Buchanan (2014) notes that the shape of the reference curve is dependent on the number of assessment layers supporting it, introducing bias towards or away from non-compliance, depending on the region of the curve. Specifically, the shape of the biological reference curve changes when it is constructed from 3, 4, or 5 annual averages instead of the 9, 12, or 15 corresponding monthly assessment layers. The reference curve bottleneck, found in the upper left region of the CFD curve, can force out of compliance segments having few interpolator grids, even if their criteria failure rates are very low. All of these issues bias towards non-compliance when using monitoring program with a low spatial density of sampling.

Furthermore, several methodological or conceptual issues exist with the current use of the CFD. First, comparison of seasonal mean chlorophyll-*a* to reference curves based on monthly sampling events represents an inconsistency; Buchanan (2014) recommends development of season and segment-specific hyperbolic reference curves. Better understanding of how reference curves are treated statistically is also warranted. Apparently, the proposed assessment curve is based on the central tendency of a family of reference curves (E. Perry, personal communication). An assessment curve that represents the central tendency of reference condition rather than the more commonly used 75th percentile, will be over-protective. Additional investigation of both reference and impacted conditions are needed to identify the appropriate statistic representing this boundary. This includes the need for better congruence between data handling methods and sample size for defining the reference curves and computing the assessment curves (E. Perry, personal communication).

Overall, the panel has the impression that the current CFD assessment approach was revised from the original concept, adopted, and implemented without sufficient additional research or exploration to understand what the method calculates and how that links back to designated use protection. If the current methodology is maintained, then this panel strongly recommends that research be undertaken to carefully document and support a refined approach.

Evaluation of Alternative Assessment Proposal

The alternative assessment proposal described in Robertson (2016) is based on the concept that the spatial median of data over a segment for each monitoring event, temporally averaged as the geometric mean over a season, produces a single value in comparison with the criterion, with no use of reference curves. The temporal frequency of attainment is replaced by a simple assessment of attainment over a six-year cycle, versus a three-year cycle of temporal exceedance frequency used in the current assessment. Robertson (2016) also recommends that a revised approach to segmentations be used. The current basis for the segmentation is salinity, which is not the best basis for factors that promote eutrophication. Robertson (2016)'s new analyses and proposed segments for chlorophyll-*a* criteria assessment procedure are focused solely on chlorophyll-*a*. Stations which are found to be uniform within a segment would be treated as a simple median of values; where non-uniformity is found, then the median of area-weighted values would be used.

The proposed methodology, though much simpler than the current method, is repeatable, transparent, and does account for the magnitude, extent, frequency and duration, albeit in a more parsimonious fashion. The Robertson (2016) report promotes that it offers several perceived advantages over the current methodology:

1) Less opportunity that data density will influence frequency of attainment and the outcome of biological reference curves, thus less potential for bias.

2) Results are more transparent and easier to explain to policymakers and the public.

3) It also links specifically to JRCC, using similar temporal scales of data aggregation.

4) Simpler treatment of inter-annual variation in attainment, with less likelihood to flip-flop in states of attainment.

That said, there are five *substantive* issues that documentation on the proposed methodology do not address or are problematic in the proposed approach:

1. While simple and straightforward, the real question is whether the current versus proposed approach produces an assessment which is more strongly linked to attainment of uses. Robertson (2016) provides arguments for why their proposed method is logistically easier to implement and easier to communicate, but no documentation of the scientific basis for an improved linkage to designated uses. Harding *et al.* (2014) utilized several pathways of impairment with inherently different temporal scales (*e.g.*, monthly to annual). Given this, what is the relationship between the chlorophyll-*a* seasonal geomean produced by the proposed approach and the attributes of the James River that represent use impairment? This seems like an area where the VA SAP should refine their analysis used in the VA DEQ (2016a) report to compare the current versus proposed assessment methodology to indicators of use impairment (HAB toxins, low DO, water clarity), focusing on how options in spatial and temporal data aggregation affect strength of relationship with low DO, HAB toxins, water clarity and PIBI.

2. Robertson (2016) note that "aggregating monitoring data too aggressively can lead to an inaccurate characterization of water quality and thus inadequate protection of resources." This is true and this panel notes that the proposed approach also represents a very aggressive aggregation of monitoring data by segment, season, and by estuarine surface area, with little documentation on rationale. A median or geomean of monitored values within a segment will have the effect of eliminating high chlorophyll-*a* values, which can be problematic if blooms are spatially patchy and peak regions of chlorophyll-*a* production are responsible for low DO, toxic HAB blooms, etc. reduce the relative importance of the high chlorophyll-*a* values in the tidal fresh regions of the James River. In order to support such aggregation, it is important to document how options in data aggregation are linked to characterization of potential adverse effects on designated uses (HABs, low DO, and water clarity).

3. Robertson (2016) proposes a six- rather than a three-year assessment window because "it is possible for two or three consecutive seasons' worth of samples to have the same skew, similar to how it is possible to roll the same number three times in a row in a dice game." The probability of rolling the same number twice in a row with a dice is 0.03 (1/6 x 1/6), and for three times in a row the probability is 0.005 (1/6 x 1/6 x 1/6). Both of those probabilities are accepted by most statisticians as an acceptable risk of being wrong. A six-year assessment window may delay for too long the management response to pursue nutrient management options. Further rationale is needed to justify why such a decision is warranted.

4. Robertson (2016) makes note that funding for monitoring is dwindling and that a more simplified approach is required to match available funding for monitoring. While we are sympathetic to this, it seems that the burden is on VA DEQ to document whether JRCC are being met; reducing monitoring effort would only weaken their case for attainment of JRCC and, presumably, delisting, regardless of the assessment approach use. If this hasn't already been done, *it seems important to determine through power analyses the minimum data density required* for the adopted assessment approach, with more specific minimum requirements on monitoring approach (*e.g.*, fixed station versus dataflow). It seems like it would make a big difference over what time frame the data are collected (*e.g.*, as would be the case in fixed stations versus dataflow) relative to the tidal cycles as well as whether the station data are spatially auto correlated. The language behind the enforcement policy could also be structured to encourage regulated parties to collect data when uncertainty is anticipated to be large.

5. A more thorough review of the basis for segmentation seems warranted. First, the recommended basis for re-segmentation is chlorophyll-*a* rather than salinity (Robertson 2016), but it seems just as important to include variables that link to the likelihood of use impairment (HABs, DO, water clarity) and collateral data that represent previously documented controls on phytoplankton productivity in providing the rationale for alternative segmentation. Second, as currently proposed, no consideration appears to be given of the spatial scale used to define an estuarine segment. Tidal excursion distance, typically 10-15 km, is the distance along an estuary where the same water sample could potentially be collected within a 6-hour window of rising or falling tide. The Robertson (2016) report questions interpolation to areas "greater than a kilometer away from where samples were actually taken," ignoring tidal excursion distances. We recommend a more complete analysis and documentation in order to support discussions of revised segmentation underpinning the chlorophyll-*a* assessment protocol.

3) Conclusions

The review panel appreciated in general the value of providing an improved scientific basis for the James River chlorophyll-*a* criteria, and in particular concurred that there is value in establishing a scientific linkage between chlorophyll-*a* levels and ecological effects related to support for designated human and aquatic life uses. This approach is in accord with national trends favoring developing numeric criteria development via explicit consideration of effects, albeit sometimes also supported by information derived from evaluation of reference sites. Moreover, we agree that as possible new data can and should be used to evaluate the scientific basis for numeric criteria and we appreciated the investments in obtaining data for this purpose.

We found that the reviewed documents left considerable uncertainty regarding the analytical approach and as a result did not provide a sound scientific rationale for developing new effects-based chlorophyll-*a* for the James River. Lacking a clear analytical approach, it was also not possible to evaluate with any confidence whether Virginia's existing chlorophyll-*a* criteria for the tidal James River are protective in the context of proposed effects-based approach. In this regard, the best available scientific rationale for the existing criteria remains that under which it was originally proposed and adopted.

At the core of our concerns are four key technical issues. These are as follows:

1) VA DEQ (2016a) did not present an adequate quantitative approach for relating chlorophyll-*a* to any particular risk of threshold exceedance and did not establish a clear methodology for classifying chlorophyll-*a* levels as "protective", "not protective" or other possible categorizations. A clear and reproducible quantitative classification methodology should be developed and applied.

2) Since there was no clear method for determining chlorophyll-*a* thresholds, subsequent arguments regarding uncertainty associated with estimated thresholds and related arguments regarding classification of existing criteria are not supported.

3) Attempts to establish a quantitative linkage between chlorophyll-*a* and acceptable risk associated with HABs, low DO or other endpoints may be hampered by inadequate articulation of policy objectives with respect to these endpoints. Based on our experience, we recommend establishing clearer policy objectives in advance (e.g., what level of exceedance of HAB thresholds is too much). Doing so will enable more effective scientific analysis of the levels of indicators (*e.g.*, chlorophyll-*a*) associated with attainment of the policy objectives.

4) Clear definition and consistent application of target spatial and temporal scales is a consistent concern that undermines the analysis. Issues of scale, including in the context of aggregation, may be affecting relationships between chlorophyll-*a* and risk endpoints. Our response to Question #1 details our own analyses that revealed the risk of bias in the way the James River

dataset is being used. Questions related to assessment methodology (Question #6) center mainly on harmonizing these scales between the analyses of chlorophyll-*a* levels associated with the risk objectives (*i.e.*, criteria development) and the assessment methodology. Our recommendation is that Virginia first clearly articulate the spatial and temporal scales associated with relating chlorophyll-*a* and risk thresholds, then devise a corresponding assessment approach to match the scales underlying the analysis.

Acknowledgements

The Review Panel would like to acknowledge the contributions of Rachel Dixon, William Ball, and Elgin Perry in the production of this review. Additional consultation with Rich Batiuk and Peter Tango were also appreciated. Data provided by VA DEQ enabled our group to carry out analyses that were informative to the issues discussed herein.

References

Bell, C. 2016. "Comments on *Empirical Relationships Linking Algal Blooms with Threats to Aquatic Life Designated Uses in the James River Estuary*" April 14, 2016. 34 p.

Buchanan, C. 2014. "From Programmatic Goals to Criteria for Phytoplankton Chlorophyll *a*". Interstate Commission of the Potomac River Basin. February 26, 2014. 48 p.

Brus, D.J., and J.J. DeGruijter. 1993. Design-based versus model-based estimates of spatial means: Theory and application in environmental soil science. Environmetrics 4(2): 123-152.

Duarte, C.M., D.J. Conley, J. Carstensen, and M. Sánchez-Camacho. 2009. Return to Neverland: shifting baselines affect eutrophication restoration targets. Estuaries and Coasts 32(1): 29-36.

Harding Jr, L.W., R.A. Batiuk, T.R. Fisher, C.L. Gallegos, T.C. Malone, W.D. Miller, M.R. Mulholland, H.W. Paerl, E.S. Perry, and P. Tango. 2014. Scientific bases for numerical chlorophyll criteria in Chesapeake Bay. Estuaries and Coasts 37(1): 134-148.

Robertson, T. 2016. "Proposed Assessment Methodology for James River Chlorophyll Criteria". Virginia Department of Environmental Quality. 22 p.

Secor, D.H., M.C. Christman, F. Curriero, D. Jasinski, E. Perry, S. Preston, K. Reckhow and M. Trice. 2006. The Cumulative Frequency Diagram Method for Determining Water Quality Attainment. Report of the Chesapeake Bay Program STAC Panel to Review of Chesapeake Bay Program Analytical Tools. Chesapeake Research Consortium. Edgewater, MD. 77 p.

Stevens Jr, D.L., and A.R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99(465): 262-278.

Suter II, G.W. 1993. Ecological risk assessment. Boca Raton FL: Lewis. 550 p.

Sutula, M. and D. Senn. 2016. Scientific basis to assess the effects of nutrients on San Francisco Bay beneficial uses. Technical Report 864. Southern California Coastal Water Research Project. Costa Mesa, CA. 63 p.

Sutula, M., R. Kudela, J.D. Hagy, L.W. Harding, D. Senn, J.E. Cloern, S. Bricker, G.M. Berg, and M. Beck. In Review. Novel Analyses of Long-Term Data Provide a Scientific Basis for Chlorophyll-a Thresholds in San Francisco Bay. Submitted to Estuarine, Coastal and Shelf Science.

U.S. EPA. 2008. Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries – 2008 Technical Support for Criteria Assessment Protocols Addendum. September 2008. EPA/903/R-08/001. Region III Chesapeake Bay Program Office. Annapolis, MD.

U.S. EPA. 2014. Next Generation Risk Assessment: Incorporation of Recent Advances in Molecular, Computational, and Systems Biology (Final Report). EPA/600/R-14/004. U.S. Environmental Protection Agency, Washington, DC.

VA DEQ 2016a. “Empirical Relationships Linking Algal Blooms with Threats to Aquatic Life Designated Uses in the James River Estuary”. A Report from the Science Advisory Panel for the James River Chlorophyll Criteria Study. April 14, 2016. 44 p.

VA DEQ 2016b. “Critical Review of the Assessment Methodology for James River Chlorophyll”. Virginia Department of Environmental Quality. 85 p.

Appendix A – STAC Review Request

STAC Independent Scientific Peer Review Panel Questions for the James River Chlorophyll *a* Criteria Re-evaluation

May 20, 2016

CBP Groups: Scientific and Technical Analysis and Reporting (STAR) Team's Criteria Assessment Protocol (CAP) Workgroup; Water Quality Goal Implementation Team (WQGIT)

CBP Contacts: Peter Tango (CAP), Lucinda Power (WQGIT)

Introduction

In 2005, the Virginia Department of Environmental Quality (DEQ) promulgated a set of tidal James River specific numerical chlorophyll *a* criteria, along with specific criteria attainment assessment procedures, into the Commonwealth's water quality standards regulations. In 2011, the Commonwealth of Virginia identified a need for additional scientific study to ensure that chlorophyll *a* criteria for the tidal James River were appropriately protective of aquatic life designated uses. The Virginia Department of Environmental Quality initiated a review of the numeric chlorophyll *a* criteria for the James and established a Science Advisory Panel to analyze the best scientific information currently available and provide recommendations as to whether the chlorophyll *a* criteria were protective of the aquatic life designated use and scientifically defensible.

Reports to be reviewed by the Panel

The Virginia Department of Environmental Quality's James River Chlorophyll a Criteria Re-evaluation Science Advisory Panel has produced a report entitled: "Empirical Relationships Linking Algal Blooms with Threats to Aquatic Life Designated Uses in the James River Estuary." The Panel report was based on an intensive 3-year research and monitoring program funded by Virginia Department of Environmental Quality as well as existing published scientific literature. This is the principal report to be peer reviewed. (44 pages)

A separate report entitled "Proposed Assessment Methodology for James River Chlorophyll Criteria", authored by Dr. Tish Robertson, Virginia Department of Environmental Quality, documents an alternative chlorophyll a criteria assessment methodology to the existing chlorophyll a criteria assessment methodology published by EPA and adopted by Virginia into their state water quality standards regulations. This is the second report to be peer reviewed. (22 pages)

Panel members are being provided with both electronic and hard copies of these two documents.

Resource Materials for the Panel

The Panel members are encouraged to consult the following resource materials for further background and insights into the process for drafting both the Empirical Relationships and Proposed Assessment Methodology reports described above.

- "From Programmatic Goals to Criteria for Phytoplankton Chlorophyll *a*", written by Dr. Claire Buchanan, Interstate Commission of the Potomac River Basin, with funding provided by Clean Water Act §106 funds from U.S. EPA Region 3, recommends consideration of a reference-based approach to criteria derivation. (44 pages)

- "Critical Review of the Assessment Methodology for James River Chlorophyll Criteria", written by Dr. Tish Robertson, Virginia Department of Environmental Quality, documents an evaluation of the existing chlorophyll a criteria assessment methodology published by EPA and adopted by Virginia into their state water quality standards regulations. (85 pages)

As part of the process for writing the Science Advisory Panel's report, panel members were encouraged to provide their comments on the draft report. Those sets of comments are listed below and provided to the Panel members in a single compiled document.

- "Observation on the Use of Arithmetic vs. Geometric Mean Chlorophyll a Targets" written by Clifton Bell, March 16, 2016. (4 pages)

- "Comment on the arithmetic vs. geometric mean interpretation of James River Chla criteria values" written by Claire Buchanan, Interstate Commission on the Potomac River Basin, April 15, 2016. (2 pages)

- "Comments on *Empirical Relationships Linking Algal Blooms with Threats to Aquatic Life Designated Uses in the James River Estuary* – version dated April 14, 2016" written by Clifton Bell, Brown and Caldwell, April 18, 2016. (34 pages)

- "Comments on *Empirical Relationships Linking Algal Blooms with Threats to Aquatic Life Designated Uses in the James River Estuary* (dated April 14, 2016)" written by Will Hunley, Hampton Roads Sanitation District, April 19, 2016. (28 pages)

- "Edits to *Empirical Relationships Linking Algal Blooms with Threats to Aquatic Life Designated Uses in the James River Estuary* – version dated April 14, 2016 Peter Tango, U.S. Geological Survey/Chesapeake Bay Program Office, April 2016. (49 pages)

- "Comments on the 3/31 draft of the James River Science Advisory Panel Chlorophyll a report" written by Peter Tango, U.S. Geological Survey/Chesapeake Bay Program Office, April 26, 2016. (2 pages)

- James River Chlorophyll SAP Survey – Peter Tango, U.S. Geological Survey/Chesapeake Bay Program Office, undated. (5 pages)

Panel members are being provided with both electronic and hard copies of the above listed documents.

Reference Documents for the Panel

The Panel members are encouraged to consult the following reference documents for further background on and documentation of the prior and more recent efforts to derive numerical chlorophyll *a* criteria for Chesapeake Bay.

- U.S. Environmental Protection Agency. 2003. *Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries*. EPA 903-R-03-002. U.S. Environmental Protection Agency, Region 3, Chesapeake Bay Program Office, Annapolis, MD. Reviewers should focus on Chapter 5 *Chlorophyll a Criteria* starting on page 101 and Chapter 6 *Recommended Implementation Procedures* starting on page 145.

- U.S. Environmental Protection Agency. 2007. *Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries. 2007 Chlorophyll Criteria Addendum*. EPA 903-R-07-005 CBP/TRS 288/07. U.S. Environmental Protection Agency, Region 3 Chesapeake Bay Program Office, Annapolis, MD.

- U.S. Environmental Protection Agency. 2008. *Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries: 2008 Technical Support for Criteria Assessment Protocols Addendum*. EPA 903-R-08-001. CBP/TRS 290-08. U.S. Environmental Protection Agency, Region 3 Chesapeake Bay Program Office, Annapolis, MD. Reviewers should focus on Chapter 5 *Chlorophyll a Criteria Assessment Procedures* starting on page 27.

- U.S. Environmental Protection Agency. 2010. *Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries: 2010 Technical Support for Criteria Assessment Protocols Addendum*. May 2010. EPA 903-R-10-002. CBP/TRS 301-10. U.S. Environmental Protection Agency, Region 3 Chesapeake Bay Program Office, Annapolis, MD. Reviewers should focus on Chapter 4 *Revisions to the Chlorophyll a Criteria Assessment Methodology* starting on page 31.

- L. W. Harding Jr., R. A. Batiuk, T. R. Fisher, C. L. Gallegos, T. C. Malone, W. D. Miller, M. R. Mulholland, H. W. Paerl, E. S. Perry and P. Tango. Scientific Bases for Numerical Chlorophyll Criteria in Chesapeake Bay. *Estuaries and Coasts*. DOI 10.1007/s12237-013-9656-6.

Electronic copies of all the above listed documents are being provided to each Panel member for their use and reference.

Expertise needed for the review team

Chesapeake Bay chlorophyll *a* criteria derivations, attainment assessments, and management applications involve the array of understandings of phytoplankton dynamics, estuarine food web dynamics, and statistical analysis and interpretation of spatially complex data. An effective review team will have members familiar with:

1. The dynamics of estuarine phytoplankton and food web dynamics and responses to changes in water quality conditions characteristic of Chesapeake Bay
2. Spatial statistics
3. Application of water quality criteria in a TMDL context.

2016 James River Chlorophyll a Criteria Re-evaluation Review Questions:

The Chesapeake Bay Program (CBP) partnership requests a scientific review that directly addresses the following questions. The review committee may also make recommendations for future work by the CBP partnership that build on the questions or are related to the scientific or management issues raised in the review. The review committee will be provided with the relevant documentation and will be given access to CBP partners to facilitate the review. The review committee will generate a written report addressing the questions for submittal to the U.S. EPA Chesapeake Bay Program Office and the Virginia DEQ. STAC independent scientific peer review comments will be considered by Virginia DEQ during the pending state rulemaking to amend the James River numeric chlorophyll *a* criteria and assessment methodology; as such, a specific response to STAC comments will not be made by Virginia DEQ. EPA will ensure the STAC peer review record includes responses to the peer review panel's comments.

1. Please comment on the scientific bases for applying a combined probability approach to derive expected frequencies of threshold exceedance as a function of mean chlorophyll *a* to determine whether attainment of these criteria would result in low rates of threshold exceedance.
2. Please comment on the approach's focus on the harmful effects of algae to derive chlorophyll criteria, rather than using reference conditions (as described in Buchanan, 2016) as an additional line of evidence.
3. Please comment on the approach for defining three categories of threshold exceedances as 'protective', 'defensible', and 'not protective' and on the approach for deciding if the categorization of these threshold exceedances are scientifically defensible. Please also comment on the general concept of applying these definitions to make the determination as to whether the existing Virginia chlorophyll *a* criteria are both protective of the aquatic life designated use and scientifically defensible.
4. Please comment on the following findings: "The results of the effects-based analysis suggest that the current criteria are defensible in that they fall below the non-protective range. In most cases, the criteria fall above the upper threshold for low risk indicating that lowering the values of the criteria may result in further improvements in water quality and phytoplankton condition. However in most cases, anticipated reductions in frequency of exceedance at attainment of the low risk threshold were small."

5. Please comment on the finding that “the criteria were found to be less protective when interpreted as geometric means, indicating that conclusions regarding protectiveness are somewhat sensitive to the methodology by which attainment of the criteria is determined.”
6. Please comment on the scientific basis for replacing the current chlorophyll *a* criteria attainment assessment procedures with the proposed alternative chlorophyll *a* criteria attainment assessment procedures.
7. Please comment on whether the scientific basis and procedures described within the Scientific Advisory Panel’s report could be used to derive new chlorophyll *a* criteria for application to other tidal habitats within Chesapeake Bay with the same salinity regimes and provide similar levels of protection of aquatic life.

Proposed Peer Review Schedule and CBP Partnership Response

The CBP partnership requests that the STAC convened independent scientific peer review panel complete their review and deliver a panel report reflecting the Panel’s collective written responses to above questions by mid-September.

The CBP partnership is committed to providing written responses to the Panel’s collective responses to above questions by November 5, 2016.